SASEC2015
Third Southern African Solar Energy Conference
11 – 13 May 2015
Kruger National Park, South Africa

MACHINE LEARNING TECHNIQUES FOR SHORT TERM SOLAR FORECASTING

Lauret P.*, David M. and Tapachès E.
*Author for correspondence
PIMENT laboratory,
University of La Reunion,
15, avenue René Cassin, 97715,
Saint-Denis,
E-mail: philippe.lauret@univ-reunion.fr

ABSTRACT

In this work, we propose a benchmarking of supervised machine learning techniques (neural networks, Gaussian processes and support vector machines) in order to forecast the Global Horizontal solar Irradiance (GHI). We also include in this benchmark a simple linear autoregressive (AR) model as well as a naive model based on persistence of the clear sky index. The models are calibrated and validated with data from Reunion Island (21.34°S ; 55.49°E). The main findings of this work are, that for hour ahead solar forecasting, the machine learning techniques slightly improve the performances exhibited by the linear AR and the persistence model. These nonlinear techniques start to outperform their simple counterparts for forecasting horizons greater than one hour.

INTRODUCTION

Solar radiation forecasting is of great importance for an efficient integration of large shares of solar energy into the electricity grid. More precisely, in order to increase the integration of solar energy into electricity grids, accurate forecasts at various horizons are needed [1]. This statement is reinforced in the case of insular grids [2]. Indeed, the intermittent character of solar energy together with the fact that the island's electricity grid is not connected, may endanger the stability of the grid and consequently the supply-demand balance. Solar forecasting may be very challenging in an insular context such as islands like Reunion island which usually experience a high spatial and temporal variability of the solar resource [3]. Due to this high variability, the insular grids can experience a drop of around 40-50% of the PV power output in minutes.

As a consequence, since the end of 2010, the French government has limited by law the total power produced by the instantaneous integration of intermittent renewables (PV and wind) into the insular grids, to 30%. Since 2011, this limit has been reached for Reunion Island. In order to assure reliable grid operation and to balance the supply and demand of energy, utilities require accurate forecasts at different granularities and for different forecast horizons. For instance, short term forecasts are needed for operational planning, switching sources or re-scheduling of means of production, programming backup, planning for reserve usage, and peak load matching [4]. Depending on the forecast horizon, different input data and forecasting models are appropriate. Statistical models with on-site measured irradiance are adequate for the very short-term time scale ranging from 5 minutes up to 6 hours [1]. Forecasts based on cloud motion vectors from satellite images [1] show a good performance for a temporal range of 30 minutes to 6 hours. For forecast horizons from about 6 hours onwards, forecasts based on Numerical Weather Prediction (NWP) models are generally more accurate [1].

In this work, we assess the performance of different models for intraday solar forecasting with a special focus on the hour ahead solar forecast. Consequently, in this work, light is shed on the use of statistical models. Indeed, the solar radiation sequence can be seen as a time series, and therefore one can build statistical models to capture the underlying random processes and predict the next values. Several statistical techniques can be employed to forecast solar radiation time series. The spectrum of methods can range, for instance, from linear models like the autoregressive (AR) model to nonlinear models like artificial Neural Networks (NNs) or Support Vector Machines (SVMs). The performances of these techniques will be compared against a simple linear model and a reference persistence model.

NOMENCLATURE

I_g	[W/m^2]	Global Horizontal solar Irradiance
I_{clsk}	[W/m^2]	Clear sky irradiance
k^*	[-]	Clear sky index

CONTEXT OF STUDY

Reunion Island exhibits a particular meteorological context dominated by a large diversity of microclimates [3]. Two main regimes of cloudiness are superposed: the clouds driven by synoptic conditions over the Indian Ocean and the orographic cloud layer generated by the local reliefs. The data used to build the models are Global Horizontal Irradiances (GHI) measured at the meteorological station of St Pierre (21°34'S ; 55°49'E, 75m a.s.l) located in the southern part of Reunion Island. Measurements are available on an hourly basis and two years of data (2012 and 2013) are used respectively for the building and the appraisal of the models. The solar irradiance is measured with a secondary standard pyranometer (CMP 11 from Kipp & Zonen). The precision of the pyranometers is ± 3.0% for the daily sum of GHI. Measurement quality is an essential asset in any solar resource forecasting study. The site of St Pierre is well maintained and has followed the radiometric techniques regarding calibration, maintenance and quality control. Each data point has been processed with SERI-QC quality control software [5].

DATA PRE-PROCESSING

In this survey, as the original solar radiation series is not stationary (daily and annual seasonalities), we used a clear sky model in an attempt to obtain a stationary hourly solar series. More precisely, we obtained a new deseasonalized series $\{k^*\}$, the so-called clear sky index series, by applying the following data transformation giving by equation (1):

$$k^* = I_g / I_{clsk} \qquad (1)$$

where I_g is the measured global irradiance and I_{clsk} is the output of a specific clear sky model.

In this work, the Bird clear sky model [6] is used to pre-process the GHI data. This clear sky model will also permit the derivation of a naive model based on the persistence of the clear sky index.

Regarding the global radiation forecasting, it is a common practice to filter out the data in order to remove night hours. In this work, we chose to apply a filtering criterion based on the solar zenith angle (SZA). Solar radiation data for which the solar zenith angle is greater than 80° have been removed.

NUMERICAL EXPERIMENTS SET-UP

The goal of this paper is to evaluate some machine learning techniques in order to predict next values of solar irradiance from only past values of the irradiance i.e. no exogenous variables are used. In other words, all forecasting methods described in this work seek to find a generic model F of the form given by equation (2):

$$\hat{k}^*(t+h) = F\left(k^*(t), k^*(t-1), \cdots, k^*(t-p)\right) \qquad (2)$$

where the sign ^ is used to identify the forecast variable and the sequence $\{k^*(t), k^*(t-1), \cdots, k^*(t-p)\}$ represents the time series of p past values of the clear sky index. The forecast horizon denoted by the letter h usually ranges from 1 hour to 6 hours (intraday solar forecasting). In our case, as mentioned above, the variable of interest is the clear sky index k^*. Given forecasts of the clear sky index, GHI forecasts can be obtained by using equation (1). All the statistical methods described in this work are supervised learning methods or data-driven approaches. As a consequence, the techniques rely on the information content embedded in the training data in order to produce forecasts on unseen data. More precisely, the models' parameters are determined with the help of n pairs of input and output examples contained in the training data. Once the model is fitted, the model can be evaluated on a test dataset. In our context, $\mathcal{D} = \{\mathbf{x}_\mathrm{i}, y_i\}_{\mathrm{i}=1}^{\mathrm{n}}$ represents the training dataset. The vector $\mathbf{x}_\mathrm{i}$ contains the p past values of the clear sky index for training and y_i refer to the corresponding value of the clear sky index for the horizon h of interest. Similarly, considering n^* test cases, we have $\mathcal{D}^* = \{\mathbf{x}_i^*, y_i^*\}_{\mathrm{i}=1}^{n^*}$ for the test dataset.

DESCRIPTION OF THE FORECASTING METHODS

Reference model (PERSistence model)

The reference model, the so-called persistence model, is given by the following equation:

$$\widehat{k^*}(t+h) = k^*(t) \qquad (3)$$

The corresponding GHI forecast can be obtained through equation (4):

$$\hat{I}_g(t+h) = I_g(t) \times I_{clsk(t+h)} / I_{clsk(t)} \qquad (4)$$

Linear model (AR model)

We also define a linear model where the future value of the clear sky index variable namely $\hat{k}^*(t+h\,)$ is assumed to be a linear combination of several past observations as shown by equation (5):

$$\widehat{k^*}(t+h) = \phi_0 + \textstyle\sum_{i=0}^{p} \phi_{i+1} k^*(t-i) + \epsilon_t \qquad (5)$$

where ϵ_t is a white noise with variance σ^2. The model's parameters are the $\{\Phi_i\}_{i=0,1,\cdots p+1}$ and p is called order of the model.

Neural network model (NN model)

A NN with d inputs, m hidden neurons and a single linear output unit defines a non-linear parameterized mapping from an input vector **x** to an output y given by the relationship:

$$y(\mathbf{x};\mathbf{w}) = \textstyle\sum_{j=1}^{m} w_j\, f\left(\sum_{i=1}^{d} w_{ji}\, x_i + b_1\right) + b_2 \qquad (6)$$

Each of the m hidden units are related to the tangent hyperbolic function $f(x) = (e^x - e^{-x})/(e^x + e^{-x})$. The parameter vector $\mathbf{w} = \left(\{w_j\}, \{w_{ji}\}, b_1, b_2\right)$, which contains a set of weights $\{w_j\}, \{w_{ji}\}$ and two biases b_1, b_2, governs the non-linear mapping and is estimated during a phase called the training or learning phase. For our application, the relationship between the output $\hat{k}^*(t+h)$ and the inputs $\{k^*(t), k^*(t-1), \cdots, k^*(t-p)\}$ has the form given by equation (7):

$$\hat{k}^*(t+h) = \sum_{j=1}^{m} w_j\, f\left(\sum_{i=0}^{p} w_{ji}\, k^*(t-i) + b_1\right) + b_2 \qquad (7)$$

Gaussian Process model (GP model)

Gaussian Processes (GPs) are a relatively recent development in non-linear modelling [7]. GPs are generally stated as a kernel-based method. Indeed, it can be shown [7] that, given *n* training samples, the prediction for an input test vector $\mathbf{x}^*$ can be seen in terms of a linear combination of *n* kernel functions; each one centered on a training point. Therefore, the forecasted clear sky index is given by equation (8):

$$\hat{k}^*(t+h) = \sum_{i=1}^{n} \alpha_i\, k_f(\mathbf{x}_i, \mathbf{x}^*) \qquad (8)$$

where $k_f(x_p, x_q) = \sigma_f^2 exp\left[\frac{-(x_p-x_q)^2}{2l^2}\right]$ denotes the squared exponential covariance function and $\mathbf{x_i}$ is the ith input training vector.

Support vector machine (SVR model)

The support vector machine (SVM) is another kernel based machine learning technique used in classification tasks and regression problems [8]. Support vector regression (SVR) is based on the application of support vector machines to regression problems. This method has been successfully applied to time series forecasting tasks [9]. As for the GPs, the prediction calculated by a SVR machine for an input test case $\mathbf{x}^*$ is given by equation (9):

$$\hat{k}^*(t+h) = \sum_{i=1}^{n} \alpha_i\, k_{rbf}(\mathbf{x}_i, \mathbf{x}^*) + b \qquad (9)$$

where k_{rbf} denotes the radial basis covariance function $k_{rbf}(x_p, x_q) = exp\left[-\gamma\left|(x_p - x_q)\right|\right]$ with hyperparameter γ. The parameter *b* (or bias parameter) is derived from the preceding equation and some specific conditions (see [10] for details).

In the case of SVR, the coefficients α_i are related to the difference of two Lagrange multipliers, which are the solutions of a quadratic programming (QP) problem [10]. Unlike GPs, it must be stressed that not all the training patterns participate to the preceding relationship. Indeed, a convenient choice of a cost function i.e. Vapnik's ε-insensitive function ([10]) in the QP problem enables to obtain a sparse solution. The latter means that only some of the coefficients α_i will be nonzero. The examples that come with non-vanishing coefficients are called Support Vectors.

RESULTS

In the realm of the solar forecasting community, the commonly used error metrics are the root mean square, mean absolute and mean bias errors (RMSE, MAE, and MBE).

For instance, the RMSE metric is given by the following equation:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(I_{g_{forecast,i}} - I_{g_{measured,i}}\right)^2} \qquad (10)$$

However, their relative counterparts (rRMSE, rMBE and rMAE) are usually preferred as the utility industry desires to understand error in relative terms rather than absolute terms [11]. Normalization is done with respect to mean ground measured irradiance of the considered period.

In this work, we chose to report the accuracy of the different forecasting methods by solely using the rRMSE. According to this definition, this error metric tends (unlike the rMAE) to be influenced by some extreme events or outliers. Nonetheless, most utility users find this metric suitable as large forecast errors result in high financial losses [1].

Hour ahead GHI forecasts

In this section, we present the results of the benchmarking study. As previously mentioned, the Bird clear sky model is used to pre-process the original GHI time series. The training of the models was operated with one year of data (2012) and the validation period covers also one year (2013). Figure 1 lists the accuracy (on the one year validation period) of the different methods in the case of hour ahead forecasts.

As shown by Figure 1, the best annual predictor is the GP model (rRMSE of 21.07%). However, it appears that it is difficult for the nonlinear methods to beat by more than 1% the persistence model (rRMSE of 21.47%).

Actually, we conducted a previous study whose goal was to analyse the sky conditions experienced by the site of St Pierre (Southern coast of Reunion island). This prior site analysis showed that the site exhibits rather stable sky conditions (mainly clear sky conditions) during a year. Therefore, it seems that for a site which experiences less variability and longer sequence of clear hours, the annual gain in rRMSE (which is the difference between the rRMSE of the persistence model and the best performer) is only of +0.4%.

However, a previous study also showed that, for a site that exhibits variable cloud situations, the machine learning techniques perform better than the persistence model. The gain in rRMSE, in this case, is in average greater than 2%.

In the next section, we make a step further by assessing the accuracy of the different methods for forecasting horizons ranging from 2 to 6 hours.

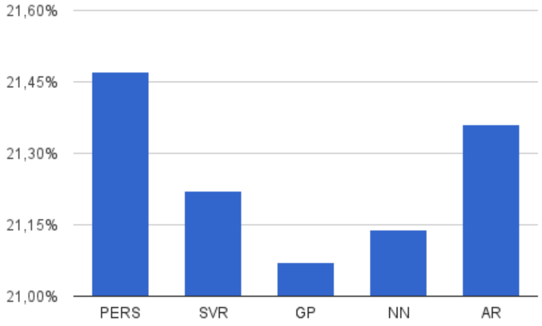

Figure 1 Relative RMSE of the different models for the hour ahead forecasts

Intra-day solar forecasting

Figure 2 shows the forecasting accuracy of the different methods for forecasting time horizons up to 6 hours. In addition to the persistence model, Figure 2 also plots the performance of another reference model. The latter, referred to as climatological mean, is independent of the forecast horizon [1]. More precisely, this model performs a constant forecast of the clear sky index that corresponds to its mean historical value. In our case, we used the average clear sky index of the year 2012 in order to forecast the clear sky index of the year 2013.

Figure 2 clearly demonstrates the better performance of the nonlinear methods over the linear AR model and persistence model when the forecast horizon increases. One may notice also that the performances of the machine learning techniques tend towards that of the climatological mean. This behavior is consistent, as these nonlinear methods tend to asymptotically model the mean of the data. As seen, it is not the case for the linear autoregressive model whose error increases with increasing forecast horizon. It should also be noted that the performance of the three nonlinear methods are practically the same. The choice of the method will depend on the skill and experience of the modeler. Nonetheless, according to our experience and as mentioned above, careful attention must be put in to the building of the NN model. Conversely, according to our experience the construction of the GP and SVR models appear to be part of a more principled framework than the NN methodology.

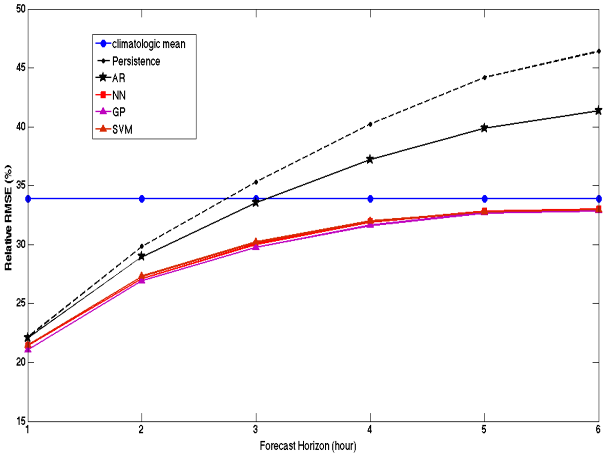

Figure 2 Relative RMSE of the different models used for the intra-day forecast

CONCLUSION

This work proposed a benchmarking of machine learning techniques for intraday solar forecasting. Popular nonlinear techniques such as neural networks, and some rather new methods such as Gaussian Processes and support vector machines were evaluated against simple methods like the autoregressive linear model and reference models like the persistence model. The main conclusion that can be drawn from this survey is that the machine learning techniques start to outperform their simple counterparts for forecasting horizons greater than one hour. For hour ahead solar forecasting, the picture is less clear and seems to depend on the sky conditions. For stable clear sky conditions (clear skies for instance), the nonlinear methods slightly improve the persistence model.

In this study, the building of the forecasting models was made solely by using past GHI measurements. The future operational availability of exogenous inputs (such as those provided by NWP models or Satellite data) will obviously improve the accuracy of the solar forecasts.

REFERENCES

[1] Lorenz E., and Heinemann D, Prediction of solar irradiance and photovoltaic power. In: Sayigh A, editor. Comprehensive Renewable Energy, Oxford:Elsevier, 2012, pp. 239-292.

[2] Diagne, M., David, M., Boland, J., Schmutz, N., and Lauret, P., Post-processing of solar irradiance forecasts from WRF model at Reunion Island, *Solar Energy*, Vol. 105, 2014, pp. 99–108.

[3] Badosa, J., Haeffelin, M., and Chepfer, H., Scales of spatial and temporal variation of solar irradiance on the tropical island of Reunion, *Solar Energy,* Vol. 88, 2013, pp. 42–56.

[4] Kostylev, V., Pavlovski, A., 2011, Solar power forecasting performance towards industry standards, *Proceedings of the 1st International Workshop on the Integration of Solar Power into Power Systems*, Aarhus, Denmark, 2011.

[5] Maxwell, E., Wilcox, S., and Rymes, M., Users manual for seri qc software, assessing the quality of solar radiation data. Report no. NREL-TP-463-5608, National Renewable Energy Laboratory,1993.

[6] Bird, R.E., and Hulstrom R.L., Simplified the Clear Sky Model for Direct and Diffuse Insolation on Horizontal Surfaces, Technical Report No. SERI/TR-642-761, Solar Energy Research Institute, 1981.

[7] Rasmussen C.E., and Williams C., Gaussian Processes for Machine Learning, MIT Press, 2006.

[8] Vapnik V., The Nature of Statistical Learning Theory, Springer, New York, 1995.

[9] Müller, K.R., , A., Ratsch, G., Scholkopf, B., Kohlmorgen, J., and V. Vapnik, V., Predicting time series with support vector machines, *Springer Lecture Notes in Computer Science,* Vol. 1327, 1997, pp. 999-1004.

[10] Smola A. and Scholkopf B., A tutorial on support vector regression, *Statistics and Computing*, Vol. 14, 2004, pp.199–222.

[11] Hoff, T.E., Perez, R., Kleissl, J., Renne, D., and Stein, J., Reporting of irradiance modeling relative prediction errors, *Progress in Photovoltaics: Research and Applications*, 2012.